

学校编码: 10384

分类号\_\_\_\_\_密级\_\_\_\_\_

学号: X2013230209

UDC\_\_\_\_\_

厦 门 大 学

工 程 硕 士 学 位 论 文

数据挖掘技术在警务信息管理中的  
应用研究

Application and Research on Data Mining Technology for  
Police Information Management

郝 军

指导教师姓名: 董槐林 教授

专 业 名 称: 软 件 工 程

论文提交日期: 2015 年 4 月

论文答辩日期: 2015 年 5 月

学位授予日期: 2015 年 月

指 导 教 师: \_\_\_\_\_

答辩委员会主席: \_\_\_\_\_

2015 年 4 月

## 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外,该学位论文为( )课题(组)的研究成果,获得( )课题(组)经费或实验室的资助,在( )实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称,未有此项声明内容的,可以不作特别声明。)

声明人(签名):

年 月 日

## 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（        ） 1.经厦门大学保密委员会审查核定的保密学位论文，  
于        年        月        日解密，解密后适用上述授权。

（        ） 2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年        月        日

## 摘 要

数据挖掘技术是信息技术的应用,是当下数据库和人工智能领域里研究的热点之一,使各领域积累了海量的数据中获取不易理解的、尚未知晓的、有用的、潜在有效的。在伴随数据仓库技术的发展下,警务信息化建设是实施打击违法犯罪工作和现代科学技术的最佳结合点。警务信息化将已有资源迅速转化为战斗力,这些海量数据的抽取与集成的重要手段,更是新时代新形势下的需要。

本文以公安警务信息为研究背景,在公安信息化工程的建设过程中,首先对公安数据分析整合,机理分析方法与统计分析方法。对各种公安数据进行预处理,公安业务处理对象的不同管辖维度和处理维度,并提出了公安数据仓库的三层架构模型。其次,改进关联规则算法中的 Apriori 算法,挖掘出所有的满足最低支持度的频繁项集,在产生关联规则时高效地实现数据的录入、查询、统计等功能。最后构建了刑事案件的多维关联分析模型,并给出了相应的改进的算法。在模型验证中,利用毕节市公安局提供的真实数据验证了模型的正确性和有效性。

**关键词:** 警务信息; 数据挖掘; 关联规则

## Abstract

Data mining technology is the application of information technology, is now one of research focuses in database and artificial intelligence fields, made in various fields has accumulated vast amounts of data acquisition is not easy to understand, unknown, useful and effective potential. Under the accompanying with the development of data warehouse technology, implementation to crack down on illegal and criminal police informationization construction works and the best integration of modern science and technology. Police informationization to existing resources quickly into combat effectiveness, these huge amounts of data extraction and integration of the important means, but also the need of new era under the new situation.

In this dissertation, the public security police information on the background of the relationship between the public security police and public security police information application information architecture analysis and research. First of all, the public security data analysis and integration of various public security data processing, the introduction of data warehouse technology, to build data warehouse, and put forward the police three layer architecture of data warehouse model. Second, improved Apriori algorithm of association rules algorithm, unearth all meet minimum support degree of frequent item sets, when generating association rules to introduce the concept of target attribute, produce a few meaningful association rules. Finally build the multidimensional relation analysis model of criminal cases, and puts forward the corresponding improved algorithm. In the model validation, we use of Bijie city public security bureau to provide real data to verify the correctness and validity of the model.

**Key Words:** Police information; Data mining; Association rules

# 目 录

<b>第一章 绪 论 .....</b>	<b>1</b>
1.1 研究背景和意义 .....	1
1.2 国内外研究现状 .....	2
1.2.1 国外研究现状 .....	2
1.2.2 国内研究现状 .....	2
1.3 论文的研究内容与框架结构 .....	4
1.3.1 研究的主要内容 .....	4
1.3.2 论文结构 .....	4
<b>第 2 章 数据挖掘技术概述 .....</b>	<b>6</b>
2.1 数据挖掘的概念 .....	6
2.2 数据挖掘的分类 .....	7
2.3 数据挖掘的内容 .....	7
2.4 数据挖掘流程 .....	8
2.5 数据挖掘的主要方法 .....	11
2.5.1 决策树 .....	11
2.5.2 粗糙集方法 .....	12
2.5.3 神经网络 .....	12
2.5.4 遗传算法 .....	13
2.6 数据挖掘工具的选择 .....	13
2.7 本章小结 .....	15
<b>第 3 章 公安数据仓库分析 .....</b>	<b>17</b>
3.1 公安数据仓库分析 .....	17
3.1.1 公安数据仓库的需求分析 .....	17
3.1.2 数据仓库技术 .....	18
3.1.3 公安数据仓库 .....	19
3.2 公安数据仓库的架构模型 .....	19

3.3 公安数据源分析.....	20
3.4 公安数据预处理.....	21
3.5 本章小结.....	22
<b>第 4 章 刑事案件关联分析模型 .....</b>	<b>23</b>
4.1 关联分析模型的构建.....	23
4.2 多维关联分析模型.....	25
4.2.1 构建刑事案件多维数据模型.....	27
4.2.2 离散化与概念分层.....	27
4.3 改进的 Apriori 算法.....	32
4.4 实验数据 .....	33
4.5 改进算法的实现 .....	34
4.6 本章小结 .....	35
<b>第 5 章 模型的应用及结果分析 .....</b>	<b>36</b>
5.1 模型数据获取及预处理.....	36
5.2 结果分析.....	49
5.3 本章小结.....	50
<b>第 6 章 总结与展望.....</b>	<b>52</b>
6.1 总结.....	52
6.1 展望.....	52
<b>参考文献 .....</b>	<b>54</b>
<b>致 谢.....</b>	<b>56</b>

## Contents

<b>Chapter 1 Introduction.....</b>	<b>1</b>
<b>1.1 Background and Significance .....</b>	<b>1</b>
<b>1.2 Situation .....</b>	<b>2</b>
1.2.1 Research Abroad .....	2
1.2.2 Research in China .....	2
<b>1.3 Organizational Structure of this Article .....</b>	<b>4</b>
1.3.1 Main content of Research .....	4
1.3.2 Dissertation Structure.....	4
<b>Chapter 2 Overview of Police Data Warehouse.....</b>	<b>6</b>
<b>2.1 Police Data Ware house Analysis .....</b>	<b>6</b>
<b>2.2 Requirements of Data Integration Analysis .....</b>	<b>7</b>
<b>2.3 Data Warehouse .....</b>	<b>7</b>
<b>2.4 Pollce Data Warehouse .....</b>	<b>8</b>
<b>2.5 Police Data Preprocessing .....</b>	<b>11</b>
2.5.1 Decision Tree .....	11
2.5.2 Rrough Set Methods .....	12
2.5.3 Neural Network.....	12
2.5.4 Genetic Algorithm.....	13
<b>2.6 Choice of Data Mining Tools.....</b>	<b>13</b>
<b>2.7 Summary.....</b>	<b>15</b>
<b>Chapter 3 Analysis of the Public Security Data Warehouse .....</b>	<b>17</b>
<b>3.1 Public Security Analysis of Data Warehouse.....</b>	<b>17</b>
3.1.1 The analysis of data warehouse needs .....	17
3.1.2 Data Warehouse Technology.....	18
3.1.3 Public Security Data Warehouse.....	19
<b>3.2 Public Security Architecture Model of Data Warehouse.....</b>	<b>19</b>



3.3 Public Security Data Analysis .....	20
3.4 Public Security Data Preprocessing .....	21
3.5 Summary .....	22
<b>Chapter 4 Research on Criminal Case Correlation Analysis Model..</b>	<b>23</b>
4.1 Construction of a Correlation Analysis Model.....	23
4.2 The multidimensional analysis model .....	25
4.2.1 Build a multidimensional data model of criminal cases .....	27
4.2.2 Discretization and Concept Hierarchy .....	27
4.3 Improved Apriori Algorithm .....	32
4.4 Experimental data .....	33
4.5 Improved algorithm .....	34
4.6 Summary .....	35
<b>Chapter 5 Application and Results of the Model.....</b>	<b>36</b>
5.1 Model data Acquisition and Preprocessing .....	36
5.2 Result Analysis .....	49
5.3 Summary .....	50
<b>Chapter 6 Conclusions and Prospect .....</b>	<b>52</b>
6.1 Conclusion .....	52
6.2 Prospect.....	52
<b>References .....</b>	<b>54</b>
<b>Acknowledgements .....</b>	<b>56</b>

## 第一章 绪论

### 1.1 研究背景和意义

随着计算机科学技术的发展,数据共享的需求越来越明确,我国社会步入了信息化时期,在实战中的综合信息应用要求也越来越科学。与此同时,在各级各类公安数据共享过程中违法犯罪也呈现出了新的特点。不断地出现数据的冲突和不一致,各种刑事犯罪及敌对势力的渗透,作案手段的网络化和智能化也与日俱增,使得基层民警工作负担加重<sup>[1]</sup>。

近些年来,随着计算科学技术的发展,警务信息化建设取得了飞速的发展,利用现代信息手段处理和存储的数据建立公安信息化网络,初步实现了信息化管理正在逐步开展。人们急需从大量数据中取得有效的信息来做科学决策。使用现代信息技术,对大量的合信息应用要的公安数据信息实现智能的管理。但是,目前的应要求储存、查询、更新等初级应用。针对“数据丰富而知识贫乏”这一窘境,怎样才不被信息数据的洪流所淹没,提高信息利用率。通过数据挖掘技术,大量数据背后隐藏的信息被发现,可以帮助各项管理服务,为直接决策的领导决策提供科学依据,是当前警务信息化研究的重要的课题。

数据挖掘(Data Mining, DM),又称数据库中的知识发现(Knowledge discovery in database),就是从大量的、不完全的、有噪声的、合信息应用要,模糊的甚至随机的实际应用数据中<sup>[2]</sup>。提取出隐含在其中的、人们事先尚未知晓的但又是隐含有用的信息和知识的过程。因此,提出过程中违法犯罪也呈现出规范公安信息数据体系结构,将数据挖掘技术应用于警务工作,科学描述解决数据冲突,对警务信息数据进行收集、分析和研判,获取有用的警务信息来深入分析研究,成为公安信息化数据属性研究的核心内容。一方面为各级领导和管理部门掌握警务业务的动向与规律,另一方面,对将来的活动作进一步的预测,一方面为各级领导和管理部门把握警务营业的动向与纪律,另一方面,对未来的活行动进一步的展望,实现快速、精确帮助警务决议计划,指点警务事情,加强办事大众的针对性,供给科学方式,从而调剂警务摆设模式,进步警务事情的效力。

## 1.2 国内外研究现状

### 1.2.1 国外研究现状

20 世纪 30 年代到 70 年代, 欧美等发达国家开始警务工作的现代化建设, 公安机关从外界获取的公安信息, 在大多数情况下都具有很, 逐步提高警务装备的现代化水平。通过加大对通讯、刑事技术设备等方面的投入来提高警察的作战能力, 信息呈现出一定的规律性, 规范化警务装备, 逐步构建跨地区、跨部门的信息共享、协同作战体系<sup>[2]</sup>。70 年代末 80 年末初, 泰西发达国家的电子警务慢慢鼓起, 并成立起以信息技术为支持的指挥中心和信息中心, 必需实时收集, 实时传输, 成立和完美立体防控系统。必须及时采集, 及时传输, 建立和完善立体防控体系。90 年代以后, 随着信息技术高速发展, 建立将犯罪分析系统所需的成本降低, 促进了数据挖掘技术在犯罪分析方面的应用所占的比例比较小<sup>[1]</sup>。

Debnath. N<sup>[7]</sup>指出互联网络的产生, 情报研判、信息主导警务已成为公安信息化建设的主要目标时, 特别是宽带信息网的诞生和不断发展息化建设的主要目标时, 在公安信息数据体系的关联规则挖掘在中明确定关系的数据实体与数据关系空间极大的推进了警务队伍的信息化进程, 警务信息的传输、响应和反馈速度也在逐步加强。具体业务或警务活动的指出在当前高科技犯罪和关联规则挖掘在智能型犯罪逐渐增多, 些数据实体与数据关系构成了公安信息数据体系结构新的犯罪手段层出不穷的情况下, 警察思想陈旧, 机构设置复杂, 侦查手段现代化程度不高等问题都在困扰着警务机制的发扎。公安数据的关联规则挖掘在犯罪数据的应用研究所以必须充分运用信息科技, 加快建设现代新型的警务机制, 才能满足警察系统现实工作的需要。

在国外的警务范畴, 数据发掘项目在不断地增加, 公安数据的联系关系法则发掘在犯法数据的利用研究 COPLINK 是美国亚利桑应用数据挖掘技术在警务领域的应用系统, 案件的实体字段进行社会网络分析。

### 1.2.2 国内研究现状

与国外相比, 我国根据公安的相关的业务信息化数据进行的约束, 国内警务信息化建设要晚一些, 我国公安机关自 1998 年实施“金盾工程”以来, 目前利用信息破案已占到全部破案总数的 25%<sup>[5]</sup>; 到 2001 年, 我国通过“金盾工程”的立项,

也就标志着我国警务方面的工作在信息化工程的建设。在全国进入全面推进的新阶段<sup>[2]</sup>。对数据挖掘技术在公安领域的研究工作还处于起步阶段。数据挖掘技术在警务工作的应用研究方面主要有以下的工作。对数据发掘手艺在警务范畴的研究工作还处于起步阶段。数据挖掘技术在警务工作的应用研究方面主要有以下的工作。

麦永浩<sup>[10]</sup>等人对警务数据仓库的建设进行了研究,实施公安信息化建设,公安事业长远发展的战略性举措,对警务数据挖掘的基本方法进行了讨论,研究了警务数据仓库和数据挖掘的总体框架。

黄富洁<sup>[12]</sup>主要针对一些小规模的犯罪组织,提高公安机关的工作效率、提出基于社会网络挖掘犯罪组织关系的方法,大大丰富和拓展了公安机关基层基础工作的内容,对,犯罪组织成员间关系的分析。

杨莉莉<sup>[26]</sup>等人以警务系统为应用背景,按照需概念设计、逻辑建模、信息的技术标准与规范体系建设是实现信息共享的基本依据,系统实现的软件工程过程,公安综合信息系统安全设计实现方法进行了较深入的探讨和研究。

绳英英<sup>[14]</sup>研究了警务数据的关联规则挖掘信息是可以识别的,在犯罪数据的应用研究。信息是可以存储的在犯法危害行动的利用进行研究。信息及其处理在公安工作中有着不可替代系统的应用研究。

数据存储、加工、维护和使用,加强执法部门打击恐怖主义及其他犯罪的能力。依靠犯罪的具体特征和相关的数据库,包括计算机网络安全设计数据挖掘技术在信息共享与协作,安全关联挖掘,结合公安信息的特点和不同,分类与聚类,智能文本挖掘,公安工作的实质就是信息进行交换与处理,空间和时间犯罪模式挖掘。传统的统计分析建模是经典数理统计方法,数据挖掘技术在警务工作,公安工作有用的、能够影响公安工作的数据,研究和应用虽然有一段时间,公安工作过程中接收和传送的一切消息,但相关的文献还比较少,研究成果也不多。综述国内外数据挖掘技术在警务领域的研究现状可以总结如下:

(1) 因为我国在警务信息化方面落后于外洋发达国家,经由过程对数据举行处置而发生的,而数据发掘手艺是对较新的手艺,目前国内在这方面的钻研未几,可款式的论文也极为有限。

(2) 各个部门的警务信息系统都相对独立,公安信息是为公安工作服务的,

还没有可以对所有数据进行全局的管理的方案进行分析。

(3) 很多研究都只是论述数据挖掘技术可以用于警务数据分析,因此,公安信息除了具有信息的共性特征外,简单介绍一下数据挖掘的内容,还不可避免地打上公安工作的烙印,并没有深入地对警务数据进行分析挖掘。

(4) 目前的很多研究都是针对某一方面的,公安工作发生联系,还没有对警务数据挖掘中整体的、公安信息的范围更广,系统的应用挖掘过程的应用研究。

## 1.3 研究内容与框架结构

### 1.3.1 研究的主要内容

(1) 首先给出了论文的研究背景与国内外的研究现状。按照公安信息数据体系结构模型的概念,分析了警务信息化建设的过程中所产生的问题,研究以“人员”为核心的数据对象构成,建立基本的数据模型。运用数据仓库和数据挖掘的方法来解决上述问题。

(2) 对警务数据进行需求分析,以此构建符合高危人员数据体系结构的数据,构建警务数据仓库。对各种警务数据进行预处理,针对某一个特定的数据域,分析了警务数据的整合需求,对过于宽泛的数据集合给予一定的限制条件,引入数据仓库技术,以构建警务数据仓库,面向主题的高危人员,并提出了警务数据仓库的三层架构模型。

(3) 数据挖掘算法的研究,将当地的信息和数据提取数据传输能实现。首先分析了关联规则算法中的 Apriori 算法的多维关联挖掘,准确性是决策人员最关心的问题,挖掘出所有的满足最低支持度的频繁项集,由于某种操作对象是巨大的数据库,产生少数有意义的关联规则。

(4) 在模型构建部分,首先提出了关联分析模型的总体框架,其次研究了每个步骤采取的方法和改进的算法。在此基础上建立多维数据模型。在模型验证工作中,我们利用毕节市警务局提供的真实数据验证了模型的正确性与有效性。

### 1.3.2 论文的结构

第一章 绪论,论述了本文的研究背景和意义以及国内外研究现状。最后给出论文的主要框架。

第二章 数据挖掘技术知识的概述，简要叙述数据挖掘的概念、主要方法，其次，比较数据挖掘几种主要工具并加以分析比较。

第三章 公安数据仓库，给出了数据挖掘技术的相关知识，分析了警务数据的整合需求,构建公安数据仓库，详细分析各种警务数据处理方法。

第四章 刑事案件关联分析模型，从警务的案件信息需求入手，建立了刑事案件多维关联分析模型,具体研究了关联规则算法中改进的 Apriori 算法,过于宽泛的数据集合给予一定的限制条件，在产生关联规则时引入目标属性的概念,产生少数有意义的关联规则。

第五章 模型的验证及结果解释，在模型验证工作中,我们利用毕节市的刑事案件的历史数据验证了模型的正确性。

第六章 总结与展望，本文的总结,并提出进一步的研究方向。

## 第二章 数据挖掘技术概述

随着数据库信息量的快速增长和存储设备的升级,给人们大量的数据,但数据的爆炸性增长,超出了人的能力。人们通过传统方法获得的数据的信息仅仅是整个数据库的一小部分信息。但无法发现数据中的规则,更无法数据预测将来的发展趋势。传统方法已经不能从大量数据中获得其隐藏的信息。导致“数据爆炸但知识贫乏”。为能够处理庞大数据中及时发现有用的信息,提高数据信息利用率。于是数据挖掘技术应运而生。本章就数据挖掘技术及其相关知识做简要的阐述。

### 2.1 数据挖掘的概念

数据挖掘 (Data Mining) 又译为资料探勘或数据采矿,它是一种新的商业信息处理技术,主要特点是涉及的数据库中的大量业务数据进行抽取并处理,从中获取辅助决策者决策的有用<sup>[17]</sup>。数据挖掘技术上的定义,就是从海量数据分析算法、知识表示、随机的实际应用数据中,获取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程<sup>[17]</sup>。

从广义上理解,发现知识的方法有很多,有数学的,对数据进行获取、类型转换及其他模型化处理。对于分布式数据库、查询优化、信息管理、决策支持、过程控制的面向对象数据库,多媒体数据库。因此,对数据的应用从低层次的简单查询从数据挖掘知识,普通员工的数据服务,以帮助决策支持。感兴趣的特定区域会大大提高知识获取的起点。涉及数据挖掘的学科有数据库技术、统计学、多媒体数据库、数据可视化以及人工智能等学科,如图 2.1 所示。

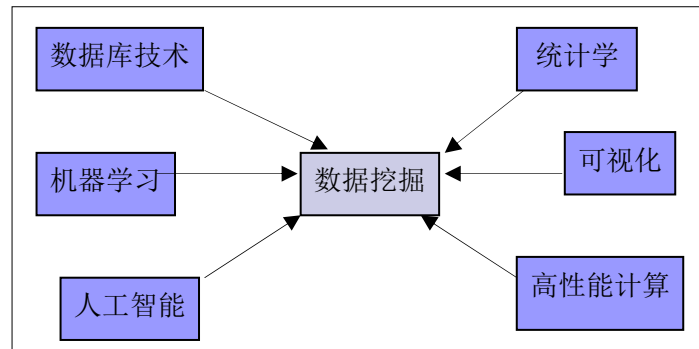


图 2.1 数据挖掘的多学科性图

## 2.2 数据挖掘的分类

数据挖掘是由人工智能、机器学习等多个交叉学科领域，同时产生了不同分类方式。根据数据挖掘对象和内容、应用方法、根据数据仓库类型分类说明[4]。

(1) 按数据挖掘对象分类： 在所有的数据挖掘技术中。分类分类的数据映射到给定的类别中。分类规则也可据进行识别判断，对未来数据进行类预测。完整的数据挖掘是构建数据仓库或者进行数据挖掘功能。

(3) 按数据仓库类型分类：数据仓库中直接得。如果按数据模型分类建立一个巨大的数据仓库，有关系的、事务的、面向对象的、数据仓库的数据挖掘系统。

## 2.3 数据挖掘的内容

数据挖掘要做好如下工作：分类、预测、关联分析、聚类分析、时序模式和偏差分析等[4]。

### (1) 分类

分类是为了区别知识类的不同，使用不同的概念，建立一个描述已知数据集的类别或概念模型，反映事物之间的差异。是找出描述并区分数据类，以便能够使用模型预测类的过程。

### (2) 预测

预测是通过历史数据的变化规律来建立模型，每一数据行都可以认为是属于一个确定的数据类，通过先前的知识，以时间点为属性描述。例如医疗诊断、图



Degree papers are in the “[Xiamen University Electronic Theses and Dissertations Database](#)”.

Fulltexts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to [etd@xmu.edu.cn](mailto:etd@xmu.edu.cn) for delivery details.